

STAT491

Summer Research Project

2006– 2007

**Advantages of Multivariate Analysis of the
Perception of Crowding on the Tongariro Crossing
Compared with Univariate Analysis**

Kathryn Baldwin

**Department of Mathematics and Statistics
University of Canterbury**

Advantages of multivariate analysis of the perception of crowding on the Tongariro Crossing compared with univariate analysis

November 2006



Kathryn Baldwin, University of Canterbury, Christchurch, New Zealand,
khb24@student.canterbury.ac.nz
Ian Westbrooke, Department of Conservation, PO Box 13049, Christchurch,
New Zealand, iwestbrooke@doc.govt.nz

Report for STAT491

Contents

	Abstract	3
1	Introduction	
1.1	Univariate Analysis	3
1.2	Multivariate Analysis	3
1.3	The Tongariro Crossing	3
2	Methods	
2.1	The Data	5
2.2	The Univariate Method	5
2.3	The Multivariate Method	6
3	Results	
3.1	Significant Results from the Univariate Analysis	6
3.2	Significant Results from the Multivariate Analysis	7
3.2.1	GAM for Description	8
3.2.2	GAM for Prediction	11
4	Discussion	13
5	References	14
	Appendix A – Survey Questionnaire	16
	Appendix B – Data screening and adjustments	19
	Appendix C - Explanatory variables and their associated categories	20
	Appendix D – Additional GAM theory	22

Abstract

Multivariate statistical analysis, in particular generalized additive models, is a very powerful approach used to investigate the significant variables that have an impact on crowding. As shown here in the 2005 visitor survey on the Tongariro Crossing. This technique provides advantages over previous univariate statistical analysis of this data. In particular it provides an objective basis to assess the crowding perceptions at different levels of walker daily numbers on the track. Generalized additive models can provide a more accurate picture of the issues currently present on the Tongariro Crossing.

1 Introduction

1.1 Univariate Analysis

The data analysis process starts with univariate analysis. Univariate analysis is the investigation into variables separately and has two purposes, firstly, to describe the data and secondly to prepare the data for multivariate modelling. The main descriptive techniques is central tendency (the three most common are mean, median and mode) and dispersion of data. Therefore, when using univariate analysis researchers are interested in the “typical” value and how far the data is dispersed from this value. Using univariate analysis to investigate the perception of crowding on Tongariro Crossing we would be interested in the occurrence of crowding, for example, the percentage of the participants who perceived the track to be crowded, and examining its relationship with each potential explanatory variable from the survey individually.

1.2 Multivariate Analysis

While univariate analysis describes the data, multivariate analysis uses a modeling approach to investigate which variables are having an influence on the data to make it appear the way it does in the context of the other potential explanatory variables. If participants of the Tongariro survey are more likely to consider the track as crowded we can use multivariate analysis to investigate which variables help explain their perception of crowding.

1.3 The Tongariro Crossing

The Tongariro National Park in the central North Island of New Zealand is the oldest national park in New Zealand. It was given to the Government as a gift on behalf of the Tuwharetoa tribe on 23 September 1887 (www.doc.govt.nz). Since then the park has grown to contain 79,598ha and, although it is still one of the smallest New Zealand National Park, it has the highest number of visitors compared with the thirteen other national parks in New Zealand (www.doc.govt.nz).

The Tongariro Crossing is a walk which usually begins at Mangatepopo in the west and finishes at Ketetahi on the northern border of the park (Figure 1). Often described

as the best one day walk in New Zealand this track is 17km in length and takes approximately 7-8 hours to complete (www.thetongarirocrossing.co.nz). The track passes over volcanic terrain and includes a variety of natural phenomena including cold mountain springs, lava flows, an active crater, steam vents, natural hot springs and emerald coloured lakes.

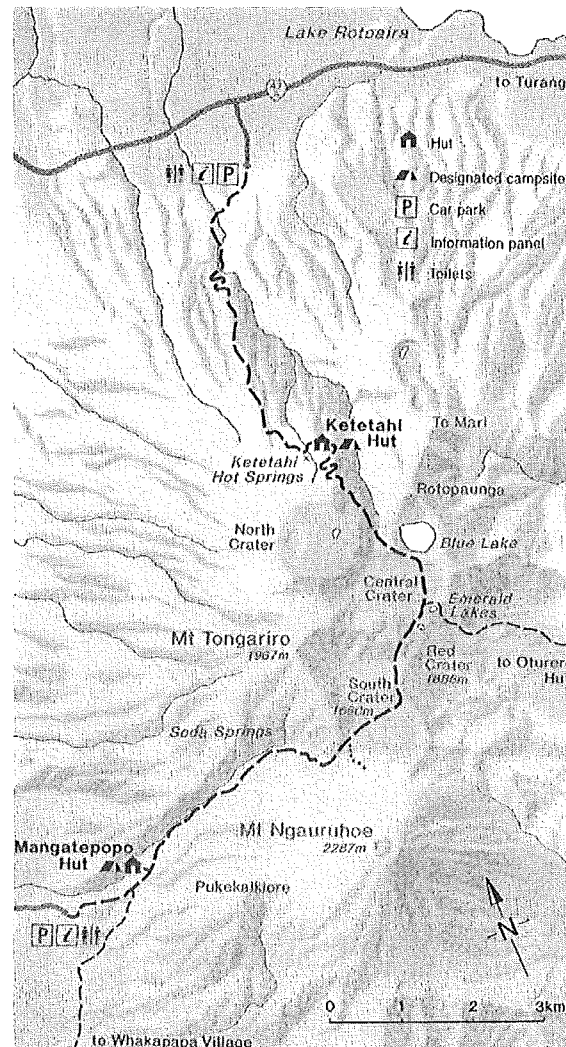


Figure 1: The Tongariro Crossing. Source: Department of Conservation

The Department of Conservation (DOC) has estimated an annual increase in the number of people using the Tongariro Crossing over the last few years (www.doc.govt.nz). If an increase continues in the future there are a number of issues that need to be considered concerning the physical, cultural and social impacts. For example safety of walkers, track degradation and toilet provisions.

The results of the 2005 walker survey are analyzed in this study to gain information about the walkers' perception of crowding on the Tongariro Crossing. In conjunction with previous analysis of this survey (Blashke, 2006) which uses univariate analysis this study applies a multivariate approach and then compares the statistical techniques.

2 Method

2.1 Data

The data is from a survey conducted in 2005 which consisted of 548 participants (Blashke, 2006). The survey form can be found in Appendix A.

Some data screening and adjustments were needed to allow statistical modeling to proceed. See Appendix B.

The focus of modeling was crowding, contrasting those who reported no crowding against any level of crowding (slightly, moderately or extremely crowded). Therefore, the response category (overall crowding) was categorized to produce a binary response – not crowded versus slightly, moderately or extremely crowded combined.

The explanatory variables used in the analysis are shown in Table 1 and the categories associated with each variable in Appendix C. These variables are characteristics of the participants walking the track and do not include any of the more subjective variables that are based on participants perception.

Age	Age of the participant
Counter	Number of walkers on the track recorded for the whole of each day only by a track counter
Country	Country of origin of the participant
Duration	The approximate length of time it took participant to complete the track
Gender	Gender of the participant
Group Size	Number of people participant walked the track with
Start Time	The approximate time participants started walking the track
Track Before	Whether participants had done the track before
Tramping Experience	The level of experience of tramping the participant had

Table 1. Definitions of Explanatory Variables used in Analysis.

2.2 Univariate Method

This method considered each variable separately to gain an overview into the perception of crowding of participants and their physical characteristics. It considers the percentage of participants in each category of the explanatory variables to see what type of walkers use the track and their perception of crowding.

2.3 Multivariate Method

We used a generalized additive model (GAM) to examine and model the relationship between crowding and the potential explanatory variables (Hastie & Tibshirani, 1990; Wood, 2006). This is an extension to standard linear regression and ANOVA techniques. Linear regression is useful where it is expected that a continuous response variable will increase (on average) a constant amount for a fixed increase in a continuous explanatory variable, while ANOVA is useful where it is expected that there will be, on average, a constant change in the response for changes from level to another of a categorical explanatory variable. Linear regression and ANOVA approaches are readily combined for a continuous response variable when there is a mix of continuous and categorical explanatory variables.

There are two issues with the crowding data that require extensions. First the response variable, crowded/not crowded is not continuous, instead taking just two values. This is dealt with by using a logistic regression (available within the GAM framework), which models the probability of being crowded (on a scale of 0 to 1). However, this involves using a mathematical transformation which makes it slightly more complicated to describe the relationship between the probability of crowding and the explanatory variable of interest. Second, it is unlikely that the relationship between crowding and the continuous explanatory variables, such as daily total numbers on the track, is a simple linear one. The advantage of the GAM is that it only needs to assume that there is a reasonably smooth relationship, and estimates the shape of the relationship from the data. Linear relationships are still available as an option for chosen variables.

Thus GAMs provide flexibility to extend regression approaches to allow for the two category response for crowding, for a mix of categorical and continuous explanatory variables. Further by using a modelling framework, it is possible to establish which set of explanatory variables best explain or predict the observed responses. This allows us to put some variables aside as having no significant explanatory power once the others are included. Also, examining the variables in concert can allow some variable to have some explanatory power which can be hidden if the explanatory variables are only considered one at a time.

It is important to note that the transformation necessary to model the crowding variables means that to represent the probability of crowding represented by the model in a clear manner requires choosing particular values of the other variables at which to carry out the calculations. We generally chose to fix the other variables at the level that was most common typical in the survey overall.

More of the technical details of GAMs are given in Appendix D.

3 Results

3.1 Significant Results from the Univariate Analysis

- Of the 540 participants in 2005 who responded to whether they thought the track was crowded 57.6% of them perceived the track as slightly, moderately or extremely crowded.
- Male participants were more likely to consider the track not crowded while female participants were more likely to rate the track as crowded.
- Participants that started the walk after 10.15am were more likely to consider the track not crowded but if they started the walk between 7.45am and 10.15am they were more likely to perceive the track as crowded.
- Three significant results were found relating to participants tramping experience. Firstly, participant with a lot of tramping experience are less likely to rate the track as not crowded compared with participants with little or moderate tramping experience. Secondly participants with a lot of tramping experience are more likely to rate the track as moderately crowded compared with participants with little or moderate tramping experience. Thirdly the participants with a lot of tramping experience are more likely to rate the track as extremely crowded. The mean number of people seen decreases with increasing tramping experience. Participants who had very little tramping experience and moderate tramping experience saw significantly more people than participants with a lot of tramping experience, yet experienced trampers were more likely to consider the track extremely crowded.
- Participants from New Zealand, North America and Continental Europe were more likely to consider the track as crowded while participants from Australia, UK and Ireland were more likely to perceive the track as not crowded as participants from Asia were equally likely to consider the track as crowded or not crowded.
- In all age categories participants were more likely to consider the track as crowded except for participants over the age of 50 years old, who were more likely to perceive the track as not crowded.
- The majority of participants considered the track to be crowded regardless of the amount of time it took them to complete the walk.
- Approximately 9% of the participants had done the walk before. Participants were more likely to consider the track as crowded regardless of whether they had done the track before.
- The majority (71%) of participants walked the track in pairs or groups of 3-4 people.

Blaschke (Blaschke, 2006) did a univariate analysis on the same data and produced similar results. His results also found that start times and nationality did not have any significant impact on the perception of overall crowding.

3.2 Significant Results from the Multivariate Analysis

The GAM procedure produces two important pieces of information. Firstly, it gives a description of each variable and the perception of crowding of the participants in each category of the variable. Secondly, it is very useful for prediction. Given the characteristics of participants the model can be used to predict the probability of the participant considering the track as crowding when there are different numbers of walkers on the track.

The GAM modeling procedure started with all the explanatory variables contained in it and then by backwards selection variables which were not significant were excluded from the model. The most appropriate model included the variables daily track count, group size, age, start time, country and tramping experience. Plots were produced showing model-based predictions of crowding for on each category of the variable.

Note that unless otherwise indicated, these predictions are made holding the other variables in the model constant at their most common or typical values, i.e. setting daily track count at 400, group size at 2, age at 25, start time at 8am, and tramping experience as moderate. This allows graphing the values or categories of each variable and the level of crowding perceived by participants. The 95% confidence intervals are represented by the dashed lines on most plots.

The results of each variable within the GAM are;

- There is a highly significant relationship between crowding and the daily track count ($p < 0.001$). If there are less than approximately 350 people on the track participants are less likely to perceive the track as crowded. While if there are more than approximately 550 people on the track participants are more likely to consider the track as crowded. Between 350 and 550 people on the track participants are equally likely to consider the track as crowded or not crowded (Figure 2).

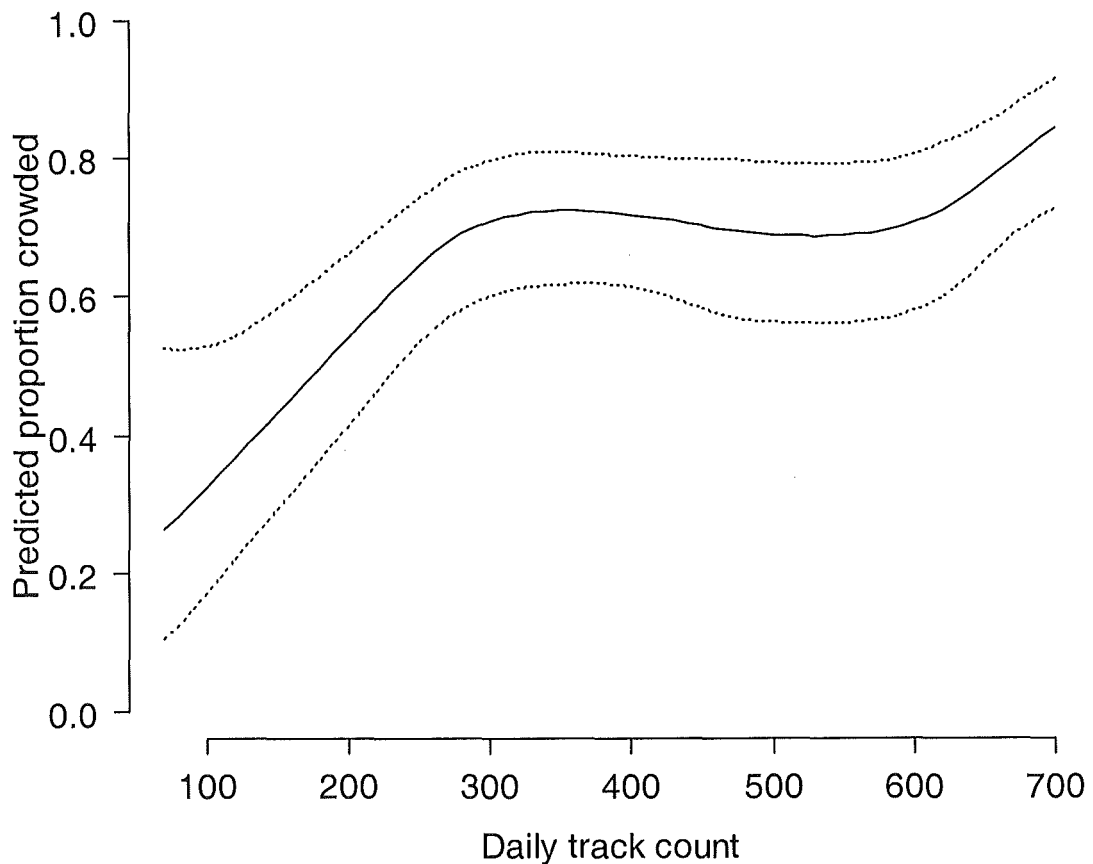


Figure 2. Predicted crowding against the count of people on the track per day, with 95 percent confidence intervals. Note that other variables in the model are at a reference level (see text)

- Age is also a highly significant factor ($p < 0.001$). As the age of participants increases they are less likely to consider the track as crowded (Figure 3)

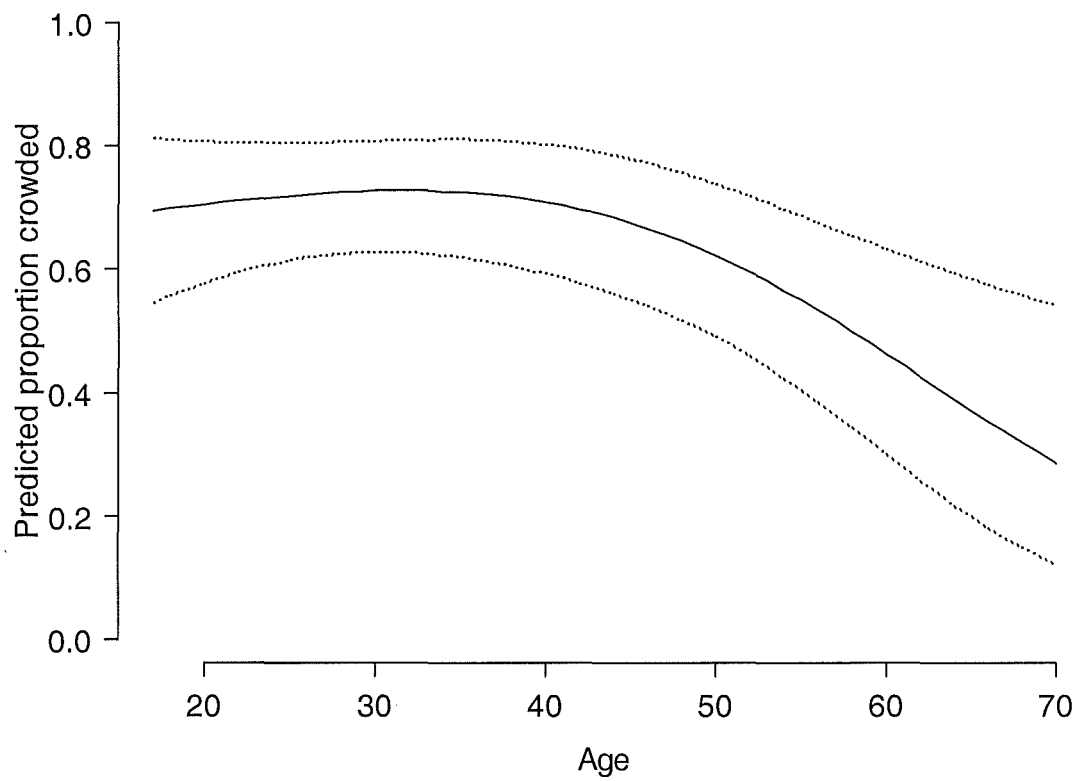


Figure 3. Predicted crowding against age, with 95 percent confidence intervals. Note that other variables in the model are at a reference level (see text)

- Start time is included based on the UBRE score indicating that the model including it is preferred, although a test comparing puts it at the very margins of significance ($p=0.08$). Participants who started before 8.15am in the morning are less likely to perceive the track as crowded compared with participants that started the track later than this time (Figure 4). Crowding reaches its highest level around 9am, but although the crowding response appears to drop from there, it is not well defined at later times, as indicated by the wide confidence intervals.

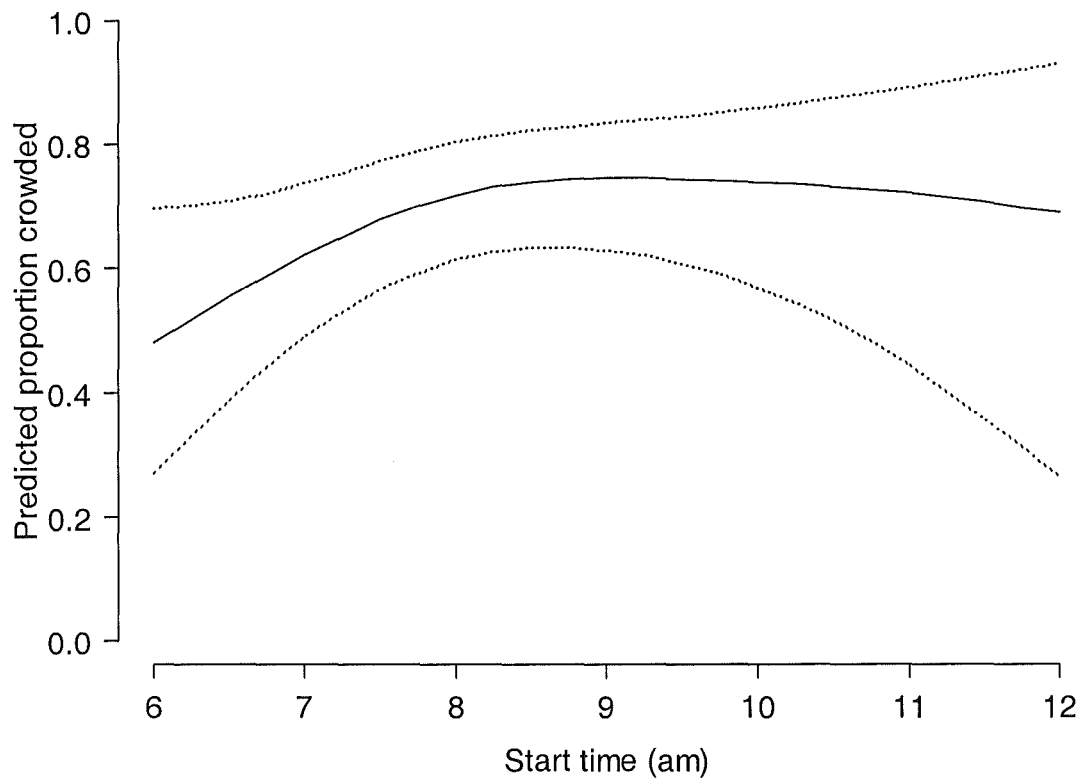


Figure 4 Predicted crowding against start time, with 95 percent confidence intervals. Note that other variables in the model are at a reference level (see text)

- The more tramping experience a participant has the more likely they will perceive the track as crowded, with overall significance for this variable being $p=0.005$. (Figure 56).

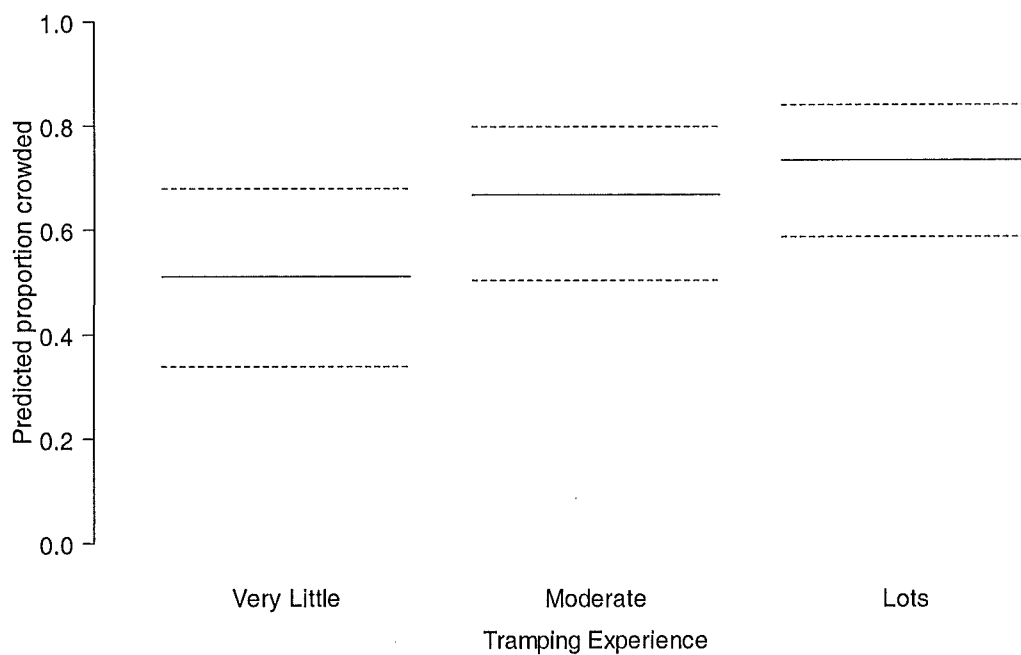


Figure 5. Predicted crowding for different levels of tramping experience, with 95 percent confidence intervals. Note that other variables in the model are at a reference level (see text)

- Group size is also a significant factor in the model for crowding ($p=0.02$) Participants that walked the track alone or in a pair had higher predicted crowding than larger groups. Crowding generally decreased with group size, except for those in groups with 9 to 20 people. Participants traveling in groups of 5-8 people perceived the track to be significantly less crowded than those people traveling alone.

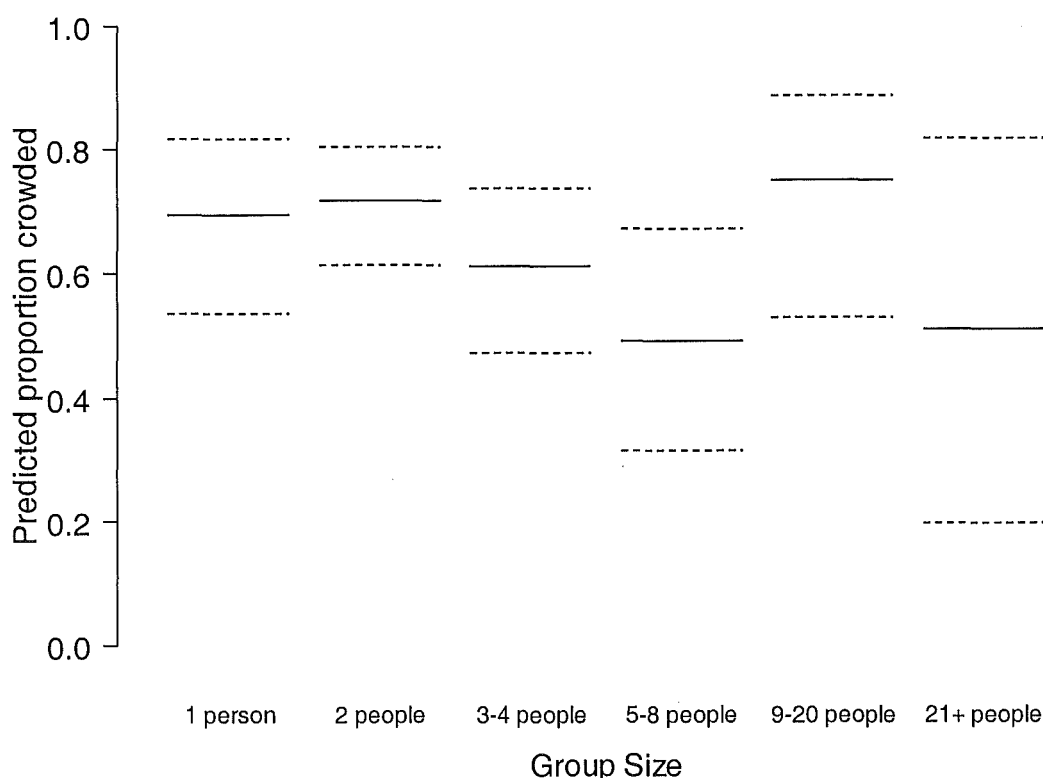


Figure 6. Predicted crowding for by groups size, with 95 percent confidence intervals. Note that other variables in the model are at a reference level (see text)

Variables not in the model

Variables considered but needed in the model were: country, duration, gender, and “track before”.

Prediction

As well as providing a means of assessing the relationship of each variable in the model with crowding, the model can be used to predict the average level of crowding expected for any given combination of the variable in the model – age, start time, tramping experience and group size. For example, given the age, start time, tramping experience, group size and the GAM can produce a probability of participants perceiving the track as crowded depending on the number of walkers on the track. This can be helpful in predicting the types of participants that would most likely

consider the track as crowded and may help in assessing measures that would reduce the impact of crowding.

4 Discussion

While the univariate analysis by Blaschke (Blaschke, 2006) and multivariate analysis are consistent in some areas for example the results of the relationship between tramping experience and perceived overall crowding, the multivariate analysis provides a much more informative and powerful analysis of overall crowding taking into consideration all the explanatory variables. The GAM model shows that country, gender, done track before and duration variables do not have a significant influence on participants' perception of overall crowding once the other variables are taken into account.

The limits of univariate analysis are quickly reached when investigating data. While univariate analysis can produce what is typical of a set of data, most researchers want to discover if this is typical of the whole population or is it more likely to occur in certain categories or groups of the data. Multivariate analysis deals with these questions that are formed from the univariate analysis and provide results that are more useful to researchers.

Generalized Additive Models are a powerful statistical analysis tool to analyze crowding. It has many advantages over univariate analysis in the Tongariro setting which include:

- It allows the response of crowding to different levels of daily use of the track to be assessed.
- It permits factors that would normally be discounted in a univariate analysis to emerge as potentially significant, for example start time.
- As GAM response curves are not restricted to linear or parabolic responses they are data driven and therefore can take any shape. This type of response curve gives the response of crowding more flexibility and a more accurate representation.
- It can be used for prediction to identify which people are more or less likely to perceive the track as crowded.

While univariate analysis is a necessity when investigating data, analysis should be taken a step further to include multivariate analysis.

References

Blaschke, P. (2006) *Establishing integrative use limits on the Tongariro Crossing, Tongariro National Park*. Department of Conservation, Wellington, New Zealand.

Hastie, T. J. and Tibshirani, R. J. (1990) *Generalized Additive Models*. Chapman & Hall, London.

Wood, S.N. (2006) *Generalized Additive Models: An Introduction with R*. Chapman & Hall, Florida.

Appendix A

Department of Conservation Crowding Survey

Interview number:

--	--	--	--

Site Name: Kaitaki Roadend

Date

Weather

Time of Day

Hello! good morning/ good afternoon. I'm doing a quick survey on behalf of the Department of Conservation – the information you give us will help us improve our management of this area. Would you have a couple of minutes to answer a few questions? IF YES: Thanks for your help. Treat questions as prompts, try not to 'read' them out.

1. First can you give us some profile information about yourself?

- Gender? ☐ Male ☐ Female • How many people are in your group? ____
- Age group? ☐ Under 20 ☐ 20-29 ☐ 30-39 ☐ 40-49 ☐ 50-59 ☐ 60+
- Where do you live? ☐ New Zealand - where? _____
☐ Overseas - what country? _____
- Have you done this track before? ☐ No ☐ Yes - if Yes, around how many times? _____
- How would you describe your experience as a walker/hiiker/supercracker?
☐ Very little experience ☐ Moderate experience ☐ Lots of experience

2. Now can you tell us a bit about your trip today?

- What time did you start? _____
- Where did you start? ☐ Mangawhaka ☐ Kaitaki ☐ Other? _____
- What time did you finish? _____
- How did you get to the start of the track?
☐ Private vehicle ☐ Campervan ☐ Bus ☐ Other (specify) _____
 If by bus - what bus company? ☐ Howards Lodge ☐ Tongariro Expeditions
☐ Alpine Scenic Tours ☐ Tongariro Track Transport ☐ Discovery Lodge
☐ Mountain Shuttle ☐ Cooe Kroll / Hukunui or Whakapapa Holiday Park

☐ Other (specify)

3. I will now ask you how you felt about the number of people on the track today.

• Did you see more or less people on the track than you expected today?

1	2	3	4	5
Saw a lot less than I expected	Saw a few less than I expected	Saw about the same as I expected	Saw a few more than I expected	Saw a lot more than I expected

• How many people do you think you saw today? (hard to estimate - ask for best guess)

• Overall, how crowded did you feel on the walk today?

<input type="checkbox"/> Not at all Crowded	<input type="checkbox"/> Slightly Crowded	<input type="checkbox"/> Moderately Crowded	<input type="checkbox"/> Extremely Crowded
--	--	--	---

• Did the number of people on the track detract from your enjoyment of the walk?

<input type="checkbox"/> Affected my enjoyment a little	<input type="checkbox"/> Affected my enjoyment moderately
<input type="checkbox"/> Did not affect my enjoyment at all	<input type="checkbox"/> Affected my enjoyment a lot
<input type="checkbox"/> Affected my enjoyment significantly	

• How many people would you be prepared to see on this walk before your enjoyment would start to diminish?

1	2	3	4	5
A lot less than today	A little less than today	About the same as today	A few more than today	A lot more than today

- If this walk was not available - would you do another in the Tongariro National Park?
☐ No ☐ Yes - If Yes, ask which walks.....

- Overall, how satisfied were you with your walk here?

1	2	3	4	5
Very	Moderately	Neutral	Moderately	Very
Dissatisfied	Dissatisfied		Satisfied	Satisfied

- Are there any improvements that could be made to this track or the facilities here?

☐ YES ☐ NO - If YES, please note specifics.....

Appendix B

Prior to statistical analysis initial data clean up was necessary.

The data was cleaned up with a few corrections and some variables removed. The column ID was removed because it was not needed to analyze the data. The comments at the end of the data set were removed and all the categorical variables which had missing responses (99) or not specified responses (98) were changed to NA (not applicable). For questions that were quantitative and where the interviewees were unsure of an exact or range of numbers, the answer was changed to NA. This made it easier for the statistical computer program R (2.2.1 A Language and Environment, 2005) to process.

Other clean-up procedures were: if a range was given in quantitative variables it was changed to a single number which was taken as the average. If the value was given as less than a specific number (e.g. < 20) then the value was changed to the number previous (e.g. <20 would be changed to 19). If a respondent gave a range greater than a certain number (e.g. > 100) this range would be changed to the number preceding it (eg. > 100 would be changed to 101).

Question 15a (coded how many people respondents thought they saw today) was left out of the data set because question 15 (actual number of people respondents thought they saw today) was already included and contained the same information.

The variables survey type and track times from the raw data were not included in the analysis. Survey type should not have any affect on whether the participants rated the track as crowded or not. The variables track before and track times appeared to overlap and contained the same information. Therefore, only the variable measuring if they had done the track previously was used. The variable which included the location in New Zealand where participants lived which was a more specific version of the country category and so only the country category was used in the analysis.

Appendix C

Variable	Categories	Coding
Age	Under 20 years	1
	20-29 years	2
	30-39 years	3
	40-49 years	4
	50-59 years	5
	60+	6
Country	New Zealand	1
	Australia	2
	UK and Ireland	3
	North America	4
	Asia	5
	Continental Europe	6
	Other	7
Duration	Less than 6 hours	1
	6 hours – 6 hours 30 minutes	2
	6 hours 31 minutes – 7 hours	3
	7 hours 1 minute – 7 hours 30 minutes	4
	7 hours 31 minutes – 8 hours	5
	8 hours 1 minute – 8 hours 30 minutes	6
	8 hours 31 minutes – 9 hours	7
	9 hours 1 minute – 9 hours 30 minutes	8
	9 hours 31 minutes – 10 hours	9
	10 hours 1 minute – 10 hours 30 minutes	10
	10 hours 31 minutes – 11 hours	11
	11 hours 1 minute – 11 hours 30 minutes	12
Gender	Male	1
	Female	2
Group Size	1 person	1
	2 people	2
	3-4 people	3
	5-8 people	4
	9-20 people	5
	21+ people	6
Start Time	5.46 – 6.15am	1
	6.16 – 6.45am	2
	6.46 – 7.15am	3
	7.16 – 7.45am	4
	7.46 – 8.15am	5
	8.16 – 8.45am	6

	8.46 – 9.15am	7
	9.16 – 9.45am	8
	9.46 – 10.15am	9
	10.16 – 10.45am	10
	10.46 – 11.15am	11
	11.16 – 11.45am	12
	11.46 – 12.15pm	13
Track Before	Yes	1
	No	2
Tramping Experience	Very Little Experience	1
	Moderate Experience	2
	Lots of Experience	3

Table 2. Explanatory variables and their associated categories

Appendix D

A Generalized Linear Model (GLM) which is a multivariate statistical technique investigates the relationship between a response variable (e.g. crowding) and many explanatory variables. It is used when there is a linear relationship between the response variable and the explanatory variables for example;

$$y = a_1x_1 + a_2x_2 + \dots + a_nx_n$$

It also allows models to be fit when the data follows distributions other than the normal distribution, ex. Binomial, Poisson and Multinomial.

A specific form of GLM is a linear logistic model which assumes that the response variable follows a binomial distribution $\text{Bin}(n(x), p(x))$ where $n(x)$ is the value of x and $p(x)$ is the probability of x . The response category (overall crowding) is re-categorized to produce a binary response. A response is stated to be 0 if a participant perceived the track not crowded and 1 if the participant thought the track was slightly, moderately or extremely crowded. Since the response variable is binary ($n(x)=1$) it follows the binomial distribution and generalized linear logistic models can be applied to the data.

The binary nature of the response variable (not crowded/crowded) allowed a binary logistic modeling approach to be applied to the data (Hastie & Tibshirani, 1990). Generalized additive models (GAM) (Wood, 2006) are an extension to logistic regression (and similar models) allowing for non-linear relationships between the response variable and some or all of the explanatory variables. For example;

$$y = a_1x_1^2 + a_2x_2^3 + \dots$$

More information about the GAM procedure can be found in Appendix D. We used package *mgcv* in R to implement the GAM models. The most appropriate GAM model was chosen as the model that minimized the Un-Biased Risk Estimator (UBRE) score. The smaller the UBRE score the better the model fits the data.

A GAM replaces the linear predictor $\sum \beta_j x_j$ by an additive predictor $\sum f_j(x_j)$. A structure of a simple logistic model structure is

$$\log \left\{ \frac{P(X)}{1 - P(X)} \right\} = \alpha + \sum_{j=1}^p f_j(X_j)$$

where $P(X) = \text{pr}(Y = 1 | X)$ and f_j is the smooth function of the explanatory variables from 1 to p . Both smoothed functions and more traditional linear functions can be combined in a GAM. The smooth function is estimated by R using penalized maximum likelihood. The number of knots, which determine how smooth the fitted additive model, is applied to each explanatory variable is a measure of how flexible the smooth is. Except where specified, we allowed the number of knots to be selected automatically by the software.

The knots for the start time and duration variables were chosen by R while the knots for group size and age were selected according to the model that minimized the AIC score.